Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2014, Article ID 240435, 7 pages
http://dx.doi.org/10.1155/2014/240435

Research Article

A Semiparametric Bivariate Probit Model for Joint Modeling of Outcomes in STEMI Patients

Francesca Ieva,[1] Giampiero Marra,[2] Anna Maria Paganoni,[3] and Rosalba Radice[4]

[1] *Department of Mathematics, Università Degli Studi di Milano, Via Saldini 50, 20133 Milano, Italy*
[2] *Department of Statistical Science, University College London, Gower Street, London WC1E 6BT, UK*
[3] *Modeling and Scientific Computing (MOX), Department of Mathematics, Politecnico di Milano, Via Bonardi 9, 20133 Milano, Italy*
[4] *Department of Economics, Mathematics and Statistics, Birkbeck, University of London, Malet Street, London WC1E 7HX, UK*

Correspondence should be addressed to Anna Maria Paganoni; anna.paganoni@polimi.it

Received 8 January 2014; Revised 25 February 2014; Accepted 10 March 2014; Published 1 April 2014

Academic Editor: Guang Wu

In this work we analyse the relationship among in-hospital mortality and a treatment effectiveness outcome in patients affected by ST-Elevation myocardial infarction. The main idea is to carry out a joint modeling of the two outcomes applying a Semiparametric Bivariate Probit Model to data arising from a clinical registry called STEMI Archive. A realistic quantification of the relationship between outcomes can be problematic for several reasons. First, latent factors associated with hospitals organization can affect the treatment efficacy and/or interact with patient's condition at admission time. Moreover, they can also directly influence the mortality outcome. Such factors can be hardly measurable. Thus, the use of classical estimation methods will clearly result in inconsistent or biased parameter estimates. Secondly, covariate-outcomes relationships can exhibit nonlinear patterns. Provided that proper statistical methods for model fitting in such framework are available, it is possible to employ a simultaneous estimation approach to account for unobservable confounders. Such a framework can also provide flexible covariate structures and model the whole conditional distribution of the response.

1. Introduction

Multiple outcomes are often used to properly characterize an effect of interest. Nevertheless, it often happens that the outcome of main interest is difficult or even impossible to measure. In general, realistic quantification of the effect of a predictor of interest on a particular response variable can be a difficult task in statistical analysis based on observational data. A solution is to control for confounders, that is, variables that are associated with both covariates and response. However, important confounders may be either unknown or too expensive to measure or not easily quantifiable (*unobservable confounders*). As pointed out in [1], this problem, which is known as *endogeneity* of the explanatory variable of interest, poses serious limitations to covariate adjustment since the use of classical techniques may yield biased and inconsistent estimates. Further issues which deserve attention are the possible presence of nonlinear covariate response relationships and how these change when considering the whole response variable distribution. There are many methods in the literature that can account for confounders. These include conditional approaches (e.g., stratification and model adjustment) and marginal approaches (e.g., matching and reweighting). For a review of these techniques see [2]. It should be noted, however, that these techniques do not account for unobserved confounding. Instrumental variable techniques are widely used for isolating the effect of a given predictor in the presence of unobserved confounding (see, among others, [3] and references therein). They are also increasingly used in epidemiological and medical studies [4]. The longest established Instrumental Variable (IV) estimators for binary outcome and treatment variables are the Generalized Method of Moment (GMM) [5], Maximum Likelihood (ML) [6], and Structural Mean Model (SMM) [7]. Among the ML estimators, the Recursive Bivariate (RB) probit model, introduced by [1], represents an effective way to estimate the effect that a binary regressor has on a binary outcome in the presence of unobservables. The semiparametric version

of BP (SBP) is an important extension since undetected nonlinearity can have severe consequences on the estimation of covariate effects; see [8], where an example of a penalized maximum likelihood fitting procedure to estimate the recursive bivariate probit model with nonlinear confounder-response relationships is proposed.

The motivating problem of this work arises from the clinical context, where multiple outcomes are often used in order to characterize the patient's status or the performances of health care service with respect to patients' management. This is a framework where unobservable confounders are very popular as well.

In such a context, during the last decade, the increased capability of data collection has made available a huge amount of information about procedures and outcomes. More and more often multiple outcomes are measured in order to characterize treatment effectiveness or to evaluate the impact of large policy initiatives. The case study considered in the following concerns patients affected by ST-Elevation Myocardial Infarction (STEMI) and admitted to any hospital of Lombardia, the Italian regional district whose capital is Milan. Data come from a clinical registry named STEMI Archive [9, 10], which is a result of a wider comprehensive project (The Strategic Program "Exploitation, integration and study of current and future health databases in Lombardia for Acute Myocardial Infarction").

In general, clinical registries and administrative databases are more and more often used nowadays to answer epidemiological enquiries (see [11–16], among others). The idea is to use information collected possibly with different purposes in order to analyze the efficacy and efficiency of the health care system on patients' outcomes. Thus, integrated health care systems for data collection measuring multiple outcomes play a fundamental role in complex clinical environments. Studies like those reported in [13–15] focus on data arising from REAL (Registro Angioplastiche dell'Emilia Romagna) and SCAAR (Swedish Coronary Angiography and Angioplasty Registry), respectively. These registries date back more than ten years and are consequently more structured and up-todated automatically with respect to STEMI Archive. Nevertheless, they are focused on the procedures (in particular, the Angioplasty and treatment of coronary in-stent restenosis), employed in the treatment of Acute Coronary Syndromes as a selection criterion for including patients in the study. Oppositely the main criterion to select eligible patients for STEMI Archive is the pathology diagnosis, being this registry more focused on imitating the classic epidemiological collections, though starting from observational data. Another difference between STEMI Archive and other registries lies in the collected information. In fact much organizational information related to hospital performances is gathered in it, in order to monitor the main process indicators in terms of time to intervention. Also concerning the goal of enhancing the integration of different sources of health information that is common to STEMI Archive, REAL, and SCAAR, STEMI Archive aims to do it also in order to automate and streamline clinicians' work flow, so that data collected once can be used multiple times for different aims. Specifically, they can serve for measuring performances of health care systems, for understanding how hospitals work, and for increasing efficacy of healthcare offer in terms of costs and patterns of care, for specific epidemiological enquires, and so on.

In STEMI Archive can be found. It consists of a clinical collection of data related to patients admitted in all hospitals of Regione Lombardia with STEMI diagnosis. As mentioned before one of the innovative contents of this survey is represented by process indicators recorded in it. They can be used to evaluate treatment times with the aim of designing a preferential therapeutic path to reperfusion in STEMI patients. In this sense, this survey represents an instrument both for epidemiological enquiries and for organizational optimization of the cardiological health care networks, quantifying the policies effects on multiple outcomes measured at patient's level. Within the data available from the STEMI Archive, there are two binary outcomes of interest for the present study: in-hospital mortality and reperfusion efficacy. The first one indicates if a patient is discharged alive from hospital. The second one indicates if a reduction greater than 70% of the ST-segment (The ECG signal can be divided into different waves and segments, delimited by some relevant points (landmarks), listed in alphabetical order starting from letter P. The P wave represents atrial depolarization; the ventricular depolarization causes the QRS complex that is followed by the ST segment. The ventricular depolarization is responsible of the T and the U waves. See [17], among others, for details.) elevation in the Electrocardiographic signal has been achieved after 1 hour from the reperfusion procedure, that is, the primary angioplasty (Percutaneous Transluminal Coronary Angioplasty or PTCA). These outcomes are clearly correlated and the interest lies in their joint modelling in order to accomplish a manyfold goal:

(1) performance evaluation (in terms of in-hospital survival) of the health care structure the patients are admitted to;

(2) quantification of the influence of procedural variables on outcomes: how does the management of the patterns of care affect the quality of life after discharge?

(3) evaluation of the relationship between outcomes, that is, success in reperfusion practices and mortality, taking advantage of the joint modeling of their dependence on categorical and continuous covariates.

The paper is then organized as follows: in Section 2 we present the recursive bivariate probit model proposed by [8], highlighting the aspects that make such a model particularly suitable for carrying out the analysis of STEMI Archive data; in Section 3 the case-study is described and results of the analysis are proposed. Section 4 contains the discussion of results and conclusions. All the analyses have been carried out using `R` software, version 3.0.2, and in particular we refer to the R-package `SemiParBIVProbit`, presented in [18].

2. Recursive Bivariate Probit Model

The bivariate probit model is a natural extension of probit regression model, where the disturbances of the two equations are assumed to be correlated in the same spirit as

the seemingly unrelated regression model [6]. The recursive version of the bivariate probit allows us to estimate the effect of interest while accounting for unobserved confounders [1, 19]. The general specification is

$$\begin{aligned} y_{1i}^{*} &= \mathbf{x}_{1i}^{\mathsf{T}}\boldsymbol{\alpha}_1 + \varepsilon_{1i}, \\ y_{2i}^{*} &= \gamma y_{1i} + \mathbf{x}_{2i}^{\mathsf{T}}\boldsymbol{\alpha}_2 + \varepsilon_{2i}, \\ i &= 1, \ldots, n, \end{aligned} \tag{1}$$

where n denotes the sample size and y_{1i}^{*} and y_{2i}^{*} are continuous latent variables (in the case of interest, the reperfusion efficacy score and propensity to die, resp.) which determine the observed binary outcomes y_{1i} (death or alive) and y_{2i} (procedure considered effective or not) through the rule $y_{vi} = 1_{\{y_{vi}^{*}>0\}}$, for $v = 1, 2$. Moreover, $\mathbf{x}_{1i}^{\mathsf{T}} = (1, \mathbf{x}_{12i}, \ldots, \mathbf{x}_{1P_1 i})$ is the ith row vector of the $n \times P_1$ model matrix $\mathbf{X}_1$, including variables related to the statistical units, in the following patients, like age, her/his total ischaemic time, killip class, and a measure of her/his ejection fraction at the entrance, and $\boldsymbol{\alpha}_1$ is a parameter vector. Similarly, $\mathbf{x}_{2i}^{\mathsf{T}}$ is the ith row vectors of the $n \times P_2$ model matrix $\mathbf{X}_2$, $\boldsymbol{\alpha}_2$ is a coefficient vector, and γ is the parameter of the endogenous binary variable y_{1i}. The error terms $(\varepsilon_{1i}, \varepsilon_{2i})$ are assumed to follow the distribution $\mathcal{N}([0, 0], [1, \rho, \rho, 1])$, where ρ is the correlation coefficient and the error variances are normalized to unity since the parameters in the model can only be identified up to a scale coefficient (e.g., [6]). To identify the parameters of the second equation in (1), it is typically assumed that the exclusion restriction on the exogenous variables holds. That is, the covariates in the first equation should contain at least one or more regressors (usually referred to as *instruments*) not included in the second equation. These regressors have to induce variation in y_{1i}, have not to directly affect y_{2i}, and have to be independent of $(\varepsilon_{1i}, \varepsilon_{2i})$ given covariates. However, as shown in [20, 21], the presence of this restriction may not be necessary to obtain consistent estimates of the model parameters.

Reference [8] proposed an extension of this model which allows for flexible functional dependence of the responses on continuous covariates: the semiparametric recursive bivariate probit model. This extension is important because the neglect of the presence of nonlinearity may have severe consequences on the estimation of covariate effects [22]. The semiparametric version of the classic bivariate probit can be written as

$$\begin{aligned} y_{1i}^{*} &= \widetilde{\mathbf{x}}_{1i}^{\mathsf{T}}\boldsymbol{\delta}_1 + \sum_{k_1=1}^{K_1} s_{1k_1}\left(\check{x}_{1k_1 i}\right) + \varepsilon_{1i}, \\ y_{2i}^{*} &= \gamma y_{1i} + \widetilde{\mathbf{x}}_{2i}^{\mathsf{T}}\boldsymbol{\delta}_2 + \sum_{k_2=1}^{K_2} s_{2k_2}\left(\check{x}_{2k_2 i}\right) + \varepsilon_{2i}, \\ i &= 1, \ldots, n, \end{aligned} \tag{2}$$

where $\widetilde{\mathbf{x}}_{1i}^{\mathsf{T}} = (1, \widetilde{x}_{12i}, \ldots, \widetilde{x}_{1Q_1 i})$ is the ith row vector of $\widetilde{\mathbf{X}}_1$, the $n \times Q_1$ model matrix for the parametric model components (such as intercept, dummy, and categorical variables), with corresponding parameter vector $\boldsymbol{\delta}_1$, and the s_{1k_1} are unknown smooth functions of the K_1 continuous covariates $\check{x}_{1k_1 i}$, which, like age and total ischaemic time, could have a nonlinear influence on the corresponding outcome. Similarly, $\check{\mathbf{x}}_{2i}^{\mathsf{T}}$ is the ith row vectors of the $n \times Q_2$ model matrix $\check{\mathbf{X}}_2$, with coefficient vector $\boldsymbol{\delta}_2$, and the s_{2k_2} are unknown smooth terms of the K_2 continuous regressors $\check{x}_{2k_2 i}$, in what follows, the measure of ejection fraction. Smooth terms are subject to identifiability constraints such as $\sum_i s_{vk_v}(\check{x}_{vk_v i}) = 0$, $v = 1, 2$, $k_v = 1, \ldots, K_v$. The smooth functions are approximated using regression splines (e.g., [23]). Here, function $s_k(\check{x}_{ki})$, where subscript v has been suppressed to avoid clutter, is given by $\sum_{j=1}^{J_k} \beta_{kj} b_{kj}(\check{x}_{ki}) = \mathbf{b}_k(\check{x}_{ki})^{\mathsf{T}}\boldsymbol{\beta}_k$, where the $b_{kj}(\check{x}_{ki})$ are known spline basis functions, with corresponding regression parameters β_{kj}, J_k is the number of spline bases, $\mathbf{b}_k(\check{x}_{ki})$ is a vector consisting of the basis functions evaluated at $\check{x}_{ki}$, that is, $\mathbf{b}_k(\check{x}_{ki})^{\mathsf{T}} = \{b_{k1}(\check{x}_{ki}), \ldots, b_{kJ_k}(\check{x}_{ki})\}$, and $\boldsymbol{\beta}_k$ is the corresponding parameter vector. Basis functions are typically chosen to have convenient mathematical properties and good numerical stability. Possible choices include B-splines, cubic regression, and thin plate regression splines (see, e.g., [24] for an overview). Based on this representation, the equations in (2) can be written as $y_{1i}^{*} = \widetilde{\mathbf{x}}_{1i}^{\mathsf{T}}\boldsymbol{\delta}_1 + \mathbf{b}_{1i}^{\mathsf{T}}\boldsymbol{\beta}_1 + \varepsilon_{1i} = \eta_{1i} + \varepsilon_{1i}$, and $y_{2i}^{*} = \gamma y_{1i} + \widetilde{\mathbf{x}}_{2i}^{\mathsf{T}}\boldsymbol{\delta}_2 + \mathbf{b}_{2i}^{\mathsf{T}}\boldsymbol{\beta}_2 + \varepsilon_{2i} = \eta_{2i} + \varepsilon_{2i}$, where, for $v = 1, 2$, $\mathbf{b}_{vi}^{\mathsf{T}} = \{\mathbf{b}_{v1}(\check{x}_{v1i})^{\mathsf{T}}, \ldots, \mathbf{b}_{vK_v}(\check{x}_{vK_v i})^{\mathsf{T}}\}$, $\boldsymbol{\beta}_v^{\mathsf{T}} = (\boldsymbol{\beta}_{v1}^{\mathsf{T}}, \ldots, \boldsymbol{\beta}_{vK_v}^{\mathsf{T}})$ and η_{vi} has the obvious definition.

2.1. Estimation and Inference. In the bivariate probit model the data identify the four possible events $(y_{1i} = 1, y_{2i} = 1)$, $(y_{1i} = 1, y_{2i} = 0)$, $(y_{1i} = 0, y_{2i} = 1)$, and $(y_{1i} = 0, y_{2i} = 0)$ with probabilities $p_{11i} = \Phi_2(\eta_{1i}, \eta_{2i}; \rho)$, $p_{10i} = \Phi(\eta_{1i}) - p_{11i}$, $p_{01i} = \Phi(\eta_{2i}) - p_{11i}$, and $p_{00i} = 1 - p_{11i} - p_{10i} - p_{01i}$, where Φ and Φ_2 are the distribution functions of a standardized univariate normal and a standardized bivariate normal with correlation ρ, respectively. Therefore, the log-likelihood function is

$$\begin{aligned} \ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \{ & y_{1i} y_{2i} \log(p_{11i}) + y_{1i}(1 - y_{2i}) \log(p_{10i}) \\ & + (1 - y_{1i}) y_{2i} \log(p_{01i}) \\ & + (1 - y_{1i})(1 - y_{2i}) \log(p_{00i}) \}, \end{aligned} \tag{3}$$

where $\boldsymbol{\theta}^{\mathsf{T}} = (\boldsymbol{\delta}_1^{\mathsf{T}}, \boldsymbol{\beta}_1^{\mathsf{T}}, \gamma, \boldsymbol{\delta}_1^{\mathsf{T}}, \boldsymbol{\beta}_2^{\mathsf{T}}, \rho)$ according to the notation introduced in the previous section. When using regression splines, to avoid overfitting, the model parameters are typically estimated by maximization of

$$\ell_p(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}) - \frac{1}{2}\boldsymbol{\beta}^{\mathsf{T}}\mathbf{S}_{\boldsymbol{\lambda}}\boldsymbol{\beta}, \tag{4}$$

where $\boldsymbol{\beta}^{\mathsf{T}} = (\boldsymbol{\beta}_1^{\mathsf{T}}, \boldsymbol{\beta}_2^{\mathsf{T}})$, $\mathbf{S}_{\boldsymbol{\lambda}} = \sum_{v=1}^{2}\sum_{k_v=1}^{K_v} \lambda_{vk_v}\mathbf{S}_{vk_v}$, and the $\mathbf{S}_{vk_v}$ are positive semidefinite known square matrices measuring the (second-order, here) roughness of the smooth terms in the model; that is, $\boldsymbol{\beta}^{\mathsf{T}}(\sum_{v=1}^{2}\sum_{k_v=1}^{K_v} \lambda_{vk_v}\mathbf{S}_{vk_v})$, $\boldsymbol{\beta} = \sum_{v=1}^{2}\sum_{k_v=1}^{K_v} \lambda_{vk_v} \int f_{vk_v}''(\check{x}_{vk_v})^2 d\check{x}_{vk_v}$. The λ_{vk_v} are smoothing

parameters controlling the trade-off between fit and smoothness. Given values for λ_{vk_v}, maximization of (4) is straightforward. However, smoothing parameter estimation has to be settled in practice. This usually involves the use of specialized numerical routines minimizing, for instance, a prediction error criterion so that the estimated smooth functions are as close as possible to the true functions. In the current context, multiple smoothing parameter estimation is achieved by minimization of the approximate unbiased risk estimator [25]. Full computational details can be found in [8].

The inferential theory for models involving penalized regression splines is not standard. This is because of the presence of penalties which undermines the use of classic asymptotic results for practical modeling. Confidence intervals (CIs) for the components in the semiparametric bivariate probit model can be constructed using the well-known Bayesian "confidence" intervals typically employed in a generalized additive model context (e.g., [26, 27]). Interval calculations are therefore based on $\boldsymbol{\theta} \mid \mathbf{y} \dot{\sim} \mathcal{N}(\widehat{\boldsymbol{\theta}}, \mathbf{V}_{\boldsymbol{\theta}})$, where $\mathbf{y}$ contains the response vectors, $\widehat{\boldsymbol{\theta}}$ is the estimate of $\boldsymbol{\theta}$, and $\mathbf{V}_{\boldsymbol{\theta}}$ represents the inverse of the penalized information matrix obtained at convergence of the algorithm. Given this result, CIs for linear and nonlinear functions of the model parameters can be easily obtained. Note that, for parametric model components, using the result above is equivalent to using classic likelihood results since such terms are not penalized. Also, there is no contradiction in fitting model (2) by penalized log-likelihood estimation and then constructing CIs adopting a Bayesian approach, and such a procedure has been employed many times in the literature (e.g., [23, 26]).

Note that in the absence of smooth functions in the model, as in (1), classic unpenalized maximum likelihood estimation can be reliably employed and traditional frequentist results used for inference. However, it is not possible to know whether smooth components are required before the analysis. In fact, using a more flexible model can help reducing the risk of misspecification due to undetected nonlinearity, which, as mentioned earlier, can have severe consequences on parameter estimation [22].

2.2. Average Treatment Effect. Since latent variables do not typically have well-defined units of measurements, parameter γ in model (1) may not be interpretable. For this reason the effect of the treatment y_{1i} on the response probability $P(y_{2i} = 1 \mid y_{1i})$ is calculated. This can be done using the average treatment effect (ATE; e.g., [3]). Given estimates for the model components, the ATE can be estimated as follows:

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\Phi_2\left(\widehat{\eta}_{2i}^{(y_{1i}=1)}, \widehat{\eta}_{1i}; \widehat{\rho}\right)}{\Phi\left(\widehat{\eta}_{1i}\right)} - \frac{\Phi_2\left(\widehat{\eta}_{2i}^{(y_{1i}=0)}, -\widehat{\eta}_{1i}; -\widehat{\rho}\right)}{1 - \Phi\left(\widehat{\eta}_{1i}\right)}, \quad (5)$$

where $\widehat{\eta}_{2i}^{(y_{1i}=r)}$ indicates the linear predictor evaluated at r equal to 1 or 0. The interpretation of this measure is straightforward; it tells us how the probability of $y_{2i} = 1$ (of dying) changes if $y_{1i} = 1$ (the procedure the patient underwent was considered effective) as compared to $y_{1i} = 1$ (the procedure was not effective).

Coefficient ρ is also of interest as it is useful to ascertain the presence of unobserved confounding (endogeneity). Specifically, ρ can be interpreted as the correlation between the unobserved confounders in the two equations (e.g., [28]). If $\rho = 0$ then ε_{1i} and ε_{2i} are uncorrelated and hence there is no problem of endogeneity. In this case, estimation of the second equation in either (1) or (2) will yield consistent parameter estimates. Moreover, if the model (1) or (2) was fitted with intercepts only, it turns out that ρ is precisely the tetrachoric correlation, that is, a Pearson correlation between two bivariate normal variables that have been observed on a dichotomous scale [29]. CIs can be obtained using the delta method (see [30]).

3. Case Study

In this section we present the analyses carried out fitting a semiparametric bivariate probit model like in (2) to the data arising from STEMI Archive, the clinical registry gathering patients admitted with ST-segment Elevation Myocardial Infarction diagnosis in any hospital of Regione Lombardia district. A complete description of this clinical registry is provided in [10], where the Archive is presented together with the motivating clinical setting. Among the most important patient information provided by this clinical registry, there are as follows:

(i) *mode of admission*, that is, if a patient reaches the hospital on her/his own or delivered by three different types of rescue units of 118 (the national toll-free number for emergencies);

(ii) *demographic features*, like age and sex;

(iii) *clinical appearance*, that is, variables describing the patient's status at admission. Among others, we focus on killip class (binary variable categorizing the severity of infarction into 0 = less severe and 1 = more severe) and Ejection Fraction (EF);

(iv) *risk factors*, like hypertension, diabetes, smoking, and chronic kidney disease (CKD);

(v) *times to treatment* (on/off hours), *times to treatment*, *times to intervention,* and all the *process indicators* concerned with pre- and in-hospital phase (Symptom Onset to Door time (OD), Door to Balloon time (DB), total ischaemic time (OB), etc.);

(vi) *clinical outcomes*, that is, in-hospital mortality and treatment efficacy (STres), quantified by a reduction of ST segment elevation in the ECG.

We focus our readings on patients who underwent primary transluminal coronary angioplasty (PTCA), the most common reperfusion procedure for acute myocardial infarctions. The population considered for the following analyses consists of 1069 statistical units.

In this application, the binary outcomes of interest are patients' in-hospital mortality and the efficacy of the reperfusion treatment they undergo. The efficacy is determined by the reduction of ST segment elevation one hour after the surgery: if the reduction is over 70% the procedure is

considered effective. Thus it is clear why the joint modelling of in-hospital mortality and reperfusion efficacy makes sense. Not only are they likely to be correlated, but also a strong clinical interest lies in quantifying the degree of correlation among these two. Moreover, reperfusion efficacy indicator is a binary variable whose values depend on the latent recovered ability of the coronary arteries to work properly. So the framework presented in [8] seems to be the proper way to address the problem of interest.

The variable selection has been carried out according to both clinical knowhow and the statistical stepwise approach, similar to what is proposed in [31–34]. Then the model (2) has been fitted to STEMI Archive data with the following specifications:

(1) for the outcome $\mathbf{y}_1^*$ (the *reperfusion efficacy*, `STres`) we retained a binary variable ($\breve{x}_{11}$) indicating if the patient reaches the hospital on her/his own (`access`) and two continuous variables ($\breve{x}_{11}$ and $\breve{x}_{12}$), being the age (`age`) of the patient and her/his total ischaemic time (`02B`, i.e., the time between the symptom onset and the PTCA procedure), respectively;

(2) for the outcome $\mathbf{y}_2^*$ (the *in-hospital mortality*, `mortality`) we retained a categorical variable ($\breve{x}_{21}$) indicating the patient's killip class (`killip`) and a continuous variable ($\breve{x}_{21}$) measuring her/his ejection fraction (`EF`) at the entrance.

Therefore, for $i = 1, \ldots, 1069$, model (2) becomes:

$$\begin{aligned} \texttt{STres}_{1i}^* &= \delta_{10} + \delta_{11} \times \texttt{access}_i \\ &\quad + \delta_{12} \times \texttt{age}_i + s_{11}\left(\texttt{02B}_{1i}\right) + \varepsilon_{1i}, \end{aligned} \tag{6a}$$

$$\begin{aligned} \texttt{Mortality}_{1i}^* &= \delta_{20} + \delta_{21} \times \texttt{STres} + \delta_{22} \times \texttt{killip} \\ &\quad + \delta_{23} \times \texttt{EF} + \varepsilon_{2i}. \end{aligned} \tag{6b}$$

Table 1 shows the estimates provided by the bivariate probit model for the mortality outcome ((6b), Table 1(b)) and the indicator of successful reperfusion therapy ((6a), Table 1(a)).

It can be noticed that all the selected covariates are significant. In particular, the treatment efficacy decreases as the age increases, as expected. The way of admission of patients delivered by 118 eases the good prognosis, too. It is worth noting that the total ischaemic time effect is nonlinear, being the smoother degrees of freedom significantly greater than one. This confirms the clinical knowhow according to which the way the delay affects the treatment efficacy is definitely nonlinear [35].

Concerning the mortality outcome, as expected, the more severe the infarction (quantified by killip class), the higher the mortality. Also a reduced ejection fraction plays the role of increasing the mortality, as it is known by clinical practice.

The estimated correlation coefficient of the recursive bivariate probit model is equal to 0.394 and it is significantly different from zero at the 5% level ($\mathrm{CI}(\rho) = (0.0637, 0.644)$), hence supporting the presence of unobserved confounders. In fact, it is usual that a lot of unexplained variability exists in complex healthcare processes where patterns of care consist of multiple phases. It derives from the variability existing at patient's level plus a variability induced by the complex process of patient's management. Then it is crucial to find correlations and to identify which procedures clinicians can act upon in order to improve the process of care.

Table 1: Parametric and smoothed coefficients' estimates obtained fitting the semiparametric bivariate PROBIT model in (2) to STEMI Archive data.

(a) Equation (6a)

	Coefficient	Estimate	Std. err.	P val.
`intercept`	δ_{10}	1.3811	0.2488	<0.0000
`access`	δ_{11}	0.2129	0.0933	0.0225
`age`	δ_{12}	−0.0088	0.0036	0.0140
	Smooth term	Edf	Est. rank	P val.
s(`02B`)	s_{11}	1.356	2	0.0041

(b) Equation (6b)

	Coefficient	Estimate	Std. err.	P val.
`intercept`	δ_{20}	1.7708	0.4656	0.0001
`STres`	δ_{21}	−1.2480	0.2109	<0.0000
`killip`	δ_{22}	0.7223	0.2544	0.0045
`EF`	δ_{23}	−0.0748	0.0119	<0.0000

Table 2: ATE estimates obtained by fitting the semiparametric bivariate probit and the naive approach, respectively.

ATE of	Estimate	CI
SBP	−2.05	(−4.88, 0.77)
AP	−1.92	(−20.63, 16.78)
Unadjusted	−4.46	(−6.32, −2.61)

Table 2 shows ATE estimates obtained using SBP. For completeness we also report the unadjusted estimate and that obtained using an additive probit model (AP). AP is a model which accounts for observed confounders but not unobserved confounders. Under this setting ATE has been estimated by fitting the equation of interest alone whereas the unadjusted estimate does not account for both observed and unobserved confounders.

It can be observed that all point estimates are negative which is consistent with the reasoning that the better the efficacy, the lower the mortality probability. However, the unadjusted ATE effect is significantly greater than those obtained using SBP and AP. This difference is due to the fact that the former estimate does not adjust for both observed and unobserved confounders and hence should be regarded with suspicion. If we account for observed confounders only, then the effect becomes strongly nonsignificant whereas the CI of ATE changes considerably if we take into account the unobservables. The modification of the CI seems to suggest that with a greater number of observations, our tenet on ATE could be confirmed. Anyway, in this case results suggest that the presence of unobserved confounders detected by the bivariate model may be regarded as variables which do not

interfere with the relationship of interest but whose presence inflates the variance of the estimates.

In this study we were concerned with the possible detrimental effect of unobserved confounders on the effect of interest (reperfusion efficacy on mortality). Based on our analysis, this does not seem to be an issue as the SBP and naive models produced similar point estimates. However, the use of a bivariate probit model may still be preferred as it may allow for more reliable inferences.

4. Discussion

It is more and more frequent in clinical practice that multiple outcomes are measured for properly characterizing an effect of interest in terms of diseases or for assessing health care policies and performances. Nevertheless, it often happens that (some) outcomes are difficult (even impossible) to be measured or that confounders are difficult to be accounted for when modeling such outcomes by means of suitable covariates. Instrumental variables are nowadays an established method for isolating the effect of a given predictor in the presence of unobserved confounding.

In this work we showed an application of a semiparametric bivariate probit model to a couple of binary outcomes representing the in-hospital mortality and an indicator of the reperfusion efficacy in patients affected by acute myocardial infarction. The efficacy is determined by the reduction of ST segment elevation one hour after the surgery: if the reduction is over 70% the procedure is considered effective. Data come from a clinical registry called STEMI Archive [9]. This case study claims for the joint modelling of the in-hospital mortality and reperfusion efficacy outcomes. It makes sense not only because they are likely to be correlated, but also because a strong clinical interest lies in quantifying the degree of correlation among these two. Moreover, reperfusion efficacy indicator is a binary variable whose values depend on the latent recovered ability of the coronary arteries to work properly. In this sense, this modelling strategy represents a step forward with respect to the results pointed out in [31] and then in [32, 36], since a joint modelling of correlated outcomes is possible, as well as parametric and nonparametric definition of the relationship between outcomes and covariates.

In fact there are many methods that can account for confounders, but many of these don not account for confounders. This approach's advantage over methods like Generalized Method of Moment (GMM) and Structural Mean Models (SMM) is twofold. First, semiparametric bivariate probit model allows for flexible functional dependence of the response variables on continuous covariates via the use of penalized regression splines. Unlike the classic parametric approach typically employed in these kinds of studies, a semiparametric specification allows us to flexibly model the effect of continuous covariates (e.g., the age of individuals) without making a priori assumptions (e.g., linearity or nonlinearity specified using quadratic or cubic polynomials). This reduces the risk of model misspecification due to undetected nonlinearity, which can have severe consequences on parameter estimation. Second, provided that the model assumptions are met, identification of the treatment effect is theoretically achieved even if an instrument is not included in the model [20, 21]. This paper does not employ those IV techniques for a variety of reasons. First, GMM and SMM do not perform well if they do not have a valid instrumental variable. More importantly, these procedures are not implemented in software-accessible code making it difficult for practitioners to use and interpret the results.

Results are strongly coherent with clinical practice. This enables a better comprehension of the disease-recovery dynamics and enables better predictions for new patients entering the study. In general, in complex processes such clinical ones where many sources of latent interactions may arise, accounting and adjusting for confounders is extremely important. This appears clearly looking at results reported in Table 2.

In general, diagnosis and management of AMI patients are difficult and may strongly benefit of the aid of statistical models that provide effective risk stratification of patients. In fact, flexible models that are able to properly profile patients adjusting for case mix and confounders are extremely of interest in the context of modern clinical practice, since the more accurate predictions and more reliable prognoses that they provide enable gaining insights of the economic burden of AMI, supporting an effective clinical decision making.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

[1] J. Heckman, "Dummy endogenous variables in a simultaneous equation system," *Econometrica*, vol. 46, pp. 931–959, 1978.

[2] P. C. Austin, "An introduction to propensity score methods for reducing the effects of confounding in observational studies," *Multivariate Behavioral Research*, vol. 46, no. 3, pp. 399–424, 2011.

[3] J. M. Wooldridge, *Econometric Analysis of Cross Section and Panel Data*, MIT Press, Cambridge, UK, 2010.

[4] D. P. Goldman, J. Bhattacharya, D. F. McCaffrey et al., "Effect of insurance on mortality in an hiv-positive population in care," *Journal of the American Statistical Association*, vol. 96, pp. 883–894, 2001.

[5] K. M. Johnston, P. Gustafson, A. R. Levy, and P. Grootendorst, "Use of instrumental variables in the analysis of generalized linear models in the presence of unmeasured confounding with applications to epidemiological research," *Statistics in Medicine*, vol. 27, no. 9, pp. 1539–1556, 2008.

[6] W. H. Greene, *Econometric Analysis*, Prentice Hall, New York, NY, USA, 2012.

[7] S. Vansteelandt and E. Goetghebeur, "Causal inference with generalized structural mean models," *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, vol. 65, no. 4, pp. 817–835, 2003.

[8] G. Marra and R. Radice, "Estimation of a semiparametric recursive bivariate probit model in the presence of endogeneity," *Canadian Journal of Statistics*, vol. 39, no. 2, pp. 259–279, 2011.

[9] Lombardia, *Determinazioni in Merito Alla Rete Per Il Trattamento Dei Pazienti Con Infarto Miocardico Con Tratto St Elevato (Stemi)*, 2009.

[10] F. Ieva, "Designing and mining a multicenter observational clinical registry concerning patients with acute coronary syndromes," in *New Diagnostic, Therapeutic and Organizational Strategies for Patients with Acute Coronary Syndromes*, N. Grieco, M. Marzegalli, and A. M. Paganoni, Eds., pp. 47–60, Springer, 2013.

[11] F. Saia, G. Piovaccari, A. Manari et al., "Patient selection to enhance the long-term benefit of first generation drug-eluting stents for coronary revascularisation procedures. Insights from a large multicentre registry," *EuroIntervention*, vol. 5, no. 1, pp. 57–66, 2009.

[12] D. Hasdai, S. Behar, L. Wallentin et al., "A prospective survey of the characteristics, treatments and outcomes of patients with acute coronary syndromes in Europe and the Mediterranean basin: The Euro Heart Survey of Acute Coronary Syndromes (Euro Heart Survey ACS)," *European Heart Journal*, vol. 23, no. 15, pp. 1190–1201, 2002.

[13] G. Campo, P. Guastaroba, A. Marzocchi et al., "Impact of copd on long-term outcome after st-segment elevation myocardial infarction receivingprimary percutaneous coronary intervention," *CHEST Journal*, vol. 144, no. 3, pp. 750–757, 2013.

[14] T. Schwalm, J. Carlsson, A. Meissner, B. Lagerqvist, and S. James, "Current treatment and outcome of coronary in-stent restenosis in sweden: a report from the swedish coronary angiography and angioplasty registry (scaar)," *EuroIntervention*, vol. 9, no. 5, pp. 564–572, 2013.

[15] T. Gudnason, G. S. Gudnadottir, B. Lagerqvist et al., "Comparison of interventional cardiology in two eu- ropean countries: a nationwide internet based registry study," *International Journal of Cardiology*, vol. 168, no. 2, pp. 1237–1242, 2013.

[16] M. Dalby, A. Bouzamondo, P. Lechat, and G. Montalescot, "Transfer for primary angioplasty versus immediate thrombolysis in acute myocardial infarction: a meta-analysis," *Circulation*, vol. 108, no. 15, pp. 1809–1814, 2003.

[17] A. E. Lindsay, *Ecg Learning Centre*, 2006.

[18] G. Marra and R. Radice, "SemiParBIVProbit: semiparametric bivariate probit modelling," R package version 3. 2-9, 2013.

[19] G. S. Maddala, *Limited Dependent and Qualitative Variables in Econometrics*, Cambridge University Press, Cambridge, UK, 1983.

[20] S. Han and E. J. Vytlacil, "Identification in a generalization of bivariate probit models with endoge- nous regressors," Working Paper, 2013.

[21] J. Wilde, "Identification of multiple equation probit models with endogenous dummy regressors," *Economics Letters*, vol. 69, no. 3, pp. 309–312, 2000.

[22] S. Chib and E. Greenberg, "Semiparametric modeling and estimation of instrumental variable models," *Journal of Computational and Graphical Statistics*, vol. 16, no. 1, pp. 86–114, 2007.

[23] S. N. Wood, *Generalized Additive Models: An Introduction with R*, Chapman & Hall/CRC, 2006.

[24] G. Marra and R. Radice, "Penalised regression splines: theory and application to medical research," *Statistical Methods in Medical Research*, vol. 19, no. 2, pp. 107–125, 2010.

[25] P. Craven and G. Wahba, "Smoothing noisy data with spline functions—estimating the correct degree of smoothing by the method of generalized cross-validation," *Numerische Mathematik*, vol. 31, no. 4, pp. 377–403, 1978.

[26] C. Gu, *Smoothing Spline ANOVA Models*, Springer, London, UK, 2002.

[27] G. Marra and S. N. Wood, "Coverage properties of confidence intervals for generalized additive model components," *Scandinavian Journal of Statistics*, vol. 39, no. 1, pp. 53–74, 2012.

[28] C. Monfardini and R. Radice, "Practitioners' corner: testing exogeneity in the bivariate probit model: A Monte Carlo Study," *Oxford Bulletin of Economics and Statistics*, vol. 70, no. 2, pp. 271–282, 2008.

[29] K. Pearson, "Mathematical contributions to the theory of evolution. VII. on the correlation of characters not quantitatively measurable," *Philosophical Transactions. Royal Society of London. Series A. Mathematical and Physical Sciences*, vol. 195, pp. 1–47, 1900.

[30] R. C. Chiburis, J. Das, and M. Lokshin, "A practical comparison of the bivariate probit and linear iv estimators," World Bank Policy Research Working Paper 5601, 2011.

[31] F. Ieva and A. M. Paganoni, "Process indicators for assessing quality of hospital care: a case study on stemi patients," *JP Journal of Biostatistics*, vol. 6, pp. 53–75, 2011.

[32] N. Grieco, F. Ieva, and A. M. Paganoni, "Performance assessment using mixed effects models: a case study on coronary patient care," *IMA Journal Management Mathematics*, vol. 23, no. 2, pp. 117–131, 2012.

[33] A. Guglielmi, F. Ieva, A. M. Paganoni, and F. Ruggeri, "A bayesian random effects model for survival probabilities after acute myocardial infarction," *Chilean Journal of Statistics*, vol. 3, pp. 1–15, 2012.

[34] A. Guglielmi, F. Ieva, A. M. Paganoni, F. Ruggeri, and J. Soriano, "Semiparametric bayesian modelingfor the classification of patients with high observed survival probabilities," *Journal of the Royal Statistical Society—Series C*, Forthcoming, 2013.

[35] B. J. Gersh, G. W. Stone, H. D. White, and D. R. Holmes Jr., "Pharmacological facilitation of primary percutaneous coronary intervention for acute myocardial infarction: is the slope of the curve the shape of the future?" *Journal of the American Medical Association*, vol. 293, no. 8, pp. 979–986, 2005.

[36] F. Ieva and A. M. Paganoni, "Multilevel models for clinical registers concerning stemi patients in a complex urban reality: a statistical analysis of momi^2 survey," *Communications in Applied and Industrial Mathematics*, vol. 1, pp. 128–147, 2010.